Interspeech 2024
1-5 September 2024, Kos, Greece

Automatic Classification of News Subjects in Broadcast News: Application to a Gender Bias Representation Analysis

Valentin Pelloin[1], *Lena Dodson*[2], *Émile Chapuis*[1], *Nicolas Hervé*[1], *David Doukhan*[1]

[1]French National Institute of Audiovisual (INA), Paris, France
[2]French Digital Communication Regulatory Authority (ARCOM), Paris, France
{vpelloin,echapuis,nherve,ddoukhan}@ina.fr, lena.dodson@arcom.fr

Abstract

This paper introduces a computational framework designed to delineate gender distribution biases in topics covered by French TV and radio news. We transcribe a dataset of 11.7k hours, broadcasted in 2023 on 21 French channels. A Large Language Model (LLM) is used in few-shot conversation mode to obtain a topic classification on those transcriptions. Using the generated LLM annotations, we explore the finetuning of a specialized smaller classification model, to reduce the computational cost. To evaluate the performances of these models, we construct and annotate a dataset of 804 dialogues. This dataset is made available free of charge for research purposes. We show that women are notably underrepresented in subjects such as sports, politics and conflicts. Conversely, on topics such as weather, commercials and health, women have more speaking time than their overall average across all subjects. We also observe representations differences between private and public service channels.
Index Terms: broadcast news, topic classification, teacher-student, LLMs, gender representation in media

1. Introduction

In April 2018, the European Parliament passed resolutions to improve gender equality in the media sector in the EU. One of its resolutions encourages "public and private media to mainstream gender equality in all their content"[1]. More generally, multiple studies have been conducted to monitor and promote equal gender representation in the media [1, 2, 3, 4].

In France, since 2016, the Regulatory Authority for Audiovisual and Digital Communication (ARCOM) is in charge (among other tasks) of collecting content reports from channels to analyze and publish reports of women representation on TV and Radio [3]. These reports are conducted on two *neutral* months which are determined beforehand, to exclude special events such as elections. For the 2023 edition, ARCOM and channels mutually chose May and October. During these months, 41 channels reported a total of 29,707 programs, categorized by program type (e.g. *Information/News*, *Documentary*, *Magazine*, or *Entertainment*). In collaboration with ARCOM, the French National Audiovisual Institute (INA) computed the speaking times using an automatic gender classification tool. In 2023, women accounted for only 34% of the total speaking time, which is significantly lower than the proportion of women in France (51.6%[2]).

However, few studies have tried to automatically estimate topic disparities based on the gender of speakers. In this paper, we want to determine, on a large-scale analysis, if men or women dominate the speaking time on specific subjects. We focus on radio and TV news broadcast, but similar analyses should also be carried out for other types of broadcast content, including fiction, documentaries, entertainment programs, etc. The contributions of this paper are as follows:

1. We describe, annotate and release a topic-classification dataset of broadcast news extracts ;
2. We evaluate three different kind of classifiers: baseline BERT models, a few-shot prompted LLM, and Teacher/Student models trained on automatically generated annotations ;
3. Using the best performing model, we process 11.7k hours of broadcast news to estimate gender representation biases in French audiovisual media, depending on content topics.

2. Related works

Over the years, automatic topic classification of news has been tackled with different techniques. Some methods involve handcrafted features, keywords counting, Bayesian modeling, or neural networks [5, 6, 7]. More recently, topic classification has been addressed by many using *finetuned* language models based on the Transformer's architecture [8, 9, 10]. Even Large Language Models (LLM) have been used in conversation mode to categorize news contents [11]. These models are often used in a few-shot configuration by prompting the model with the desired task and taxonomy. Most of these techniques have been applied for classification of written text news, however, some have been applied on automatic transcripts of audiovisual pre-segmented news subjects [9, 12].

Despite the generalization capabilities praised by some, LLMs face a significant limitation due to their computational power requirements, resulting in substantial financial costs [13]. Methods known as knowledge distillation are explored by some to project the LLMs capabilities into smaller, energy efficient models [13, 14, 15, 16, 17, 18, 19]. These techniques typically employ a large model as the *Teacher*, which generates synthetic training data for the smaller *Student* model. In some cases, the *Student* model has been found to achieve better results than the *Teacher* model [17, 19] for tasks such as Named Entity Recognition, Natural Language Generation and Understanding.

Datasets and taxonomies related to news classification include AG News, Reuters News, or AFP datasets [6, 7]. The International Press Telecommunications Council (IPTC) defines a hierarchical taxonomy of news subjects, comprising 17 top-level categories such as *weather* or *politics*. While several studies report high accuracy in automatic IPTC categorization based on text press news [7, 10, 11], approaches relying on TV news programs are scarce and have been found to be much more challenging, often requiring manual pre-segmentation of news into homogeneous topics [12].

[1]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018IP0101
[2]https://www.insee.fr/fr/statistiques/2381474

10.21437/Interspeech.2024-1854

3. Data description

We choose to conduct our analysis on French broadcast news, following the scope of ARCOM 2023 channels reports. Using available reports, we filtered programs declared on the *Information/News* type, or broadcasted from 24/7 news cycle channels. We preprocess this corpus by transcribing it using an ASR model and merging utterance segments into larger chunks for additional context. Next, we manually annotate a subsample of this dataset (804 dialogues with a total duration of 03h44m).

3.1. Preprocessing

Our first preprocessing step is to transcribe the ARCOM-declared programs using the Whisper model [20], specifically, the model released under *whisper-large-v3* name. We use the WhisperX [21] implementation, offering a speedup of 11.8x over the base implementation. ASR evaluations were conducted using an internal French 10h TV news dataset broadcasted from 2019 to 2023, which shares similar properties with our current material, resulting in a WER of 10.66%.

A total of 11.7k hours of audiovisual data are transcribed on NVIDIA RTX 2080 Ti GPU cards, with an average speed of 228s per hour of speech. 56.3% of the transcribed hours are from 24/7 news cycle TV channels, with the remaining from other channels, under the ARCOM *Information/News* type.

Broadcast news feeds are segmented into utterances using segment-level timecodes obtained from WhisperX. On average, these segments have a duration of 4 seconds, which would contain too little information to be efficiently classified into topics. A simple heuristic is implemented to group utterance segments into longer units, which will be referred to as *dialogues*. We sequentially merge segments that had a temporal gap with other segments of less than 10s, and a total duration of less than 60s. As a result, dialogues have a duration of 17 seconds. Our heuristic does not try to ensure that all dialogues contains only one single news subject. This could impact our follow-up analysis, and we therefore tried to strike a balance between too little context and too many unrelated topics inside a single dialogue.

3.2. Annotation guidelines

We define a set of 18 topic categories inspired from the IPTC taxonomy. Due to the special nature of TV and radio extracts, we decide to add the classes *commercial* for adverts, as well as an *other* category for everything that cannot not be classified in the other 17 categories. For instance, dialogues that are essentially composed of greetings fall in this *other* category. After browsing some examples manually, we decided not to include the *human interest* category, as very few dialogues could be labeled as it. The list of categories is shown in Table 1 along with their respective proportions on the annotated dataset.

Annotators are provided with a short description of the category. They are asked to label a dialogue with all relevant categories (multilabel). The order of categories is not considered important. While annotators are encouraged to share their hesitations with others, they are instructed to rely on their own judgments in case of disagreement with other annotators.

Along with the topic categories, we ask annotators to label each dialogue with the scope significance of the news relative to France (local, national, european, and international) ; whether the subject is about the Russo-Ukrainian war (6.3% of dialogues) and/or the Israel–Hamas war (16.9%). Lastly, as the dialogues are automatically assembled from speech utterances without any topic segmentation, we ask annotators to indicate when dialogues are perceived as a succession of different subjects (8.2%).

Table 1: *List of the available topics, proportions in the 804 annotated dialogues, durations, and Krippendorff's alpha (α) agreement score. Percentages and durations are shown with averaged agreements between annotators.*

Topic	#Dial.	Duration	α
religion, belief	1.0%	02m 23s	0.50
science, technology	2.3%	05m 17s	0.36
education	2.8%	06m 53s	0.52
disaster, accident	2.9%	06m 41s	0.72
labour	4.1%	09m 26s	0.49
weather	4.4%	10m 05s	0.87
health	4.5%	10m 33s	0.68
other	4.6%	07m 42s	0.43
environmental issue	6.3%	15m 06s	0.69
sport	6.5%	14m 06s	0.95
lifestyle, leisure	6.8%	14m 43s	0.57
social issue	7.7%	17m 50s	0.25
economy, business, finance	7.7%	18m 37s	0.74
commercial	9.4%	17m 26s	0.93
arts, culture, entertainment	9.8%	22m 41s	0.70
crime, law, justice	14.5%	33m 17s	0.67
politics	16.2%	39m 57s	0.62
unrest, conflicts, war	27.2%	1h 05m 51s	0.82
total	100.0%	03h 44m 44s	0.60

3.3. Annotation campaign

A total of 804 dialogues are randomly selected for manual annotation: 402 from 24/7 news cycle channels (France Info TV, CNews, LCI, France 24, BFMTV), along with 402 dialogues declared under the *Information/News* category from 7 remaining channels (RTL, TF1, M6, RMC, France Info Radio, Europe 1, France 2). Corpus annotation is realized by 3 male speech analysis or NLP researchers (aged 41, 35 and 26) and a female professional specialized in gender representation issues in media (aged 28). To mitigate the subjectivity and complexity of the task, each dialogue is annotated by two annotator, in a mix-and-match way, to compute agreements among all annotators, resulting in 402 dialogues annotated per user. Annotators spent around 7 hours for this task. For each dialogue, annotators were able to read the transcription of Whisper, along with the audio and video (for TV programs).

The average amount of identified topics per annotated dialogues depending on the annotator varies between 2.55 and 2.62. We compute the inter-annotator agreement using Krippendorff's alpha [22] for the topic categories. Results are shown in Table 1 (higher is better). The global alpha is 0.60, indicating diverse annotation strategies, which is consistent with the subjectivity and complexity of this annotation task, together with the dialogue segmentation strategy that results in short out-of-context excerpts. Some classes are much more ambiguous, like *social issue* with an alpha of only 0.25. This class could have benefited from a clearer redefinition of its description and range. Conversely, classes like *commercial* and *sport* are (almost) unquestionably labeled by the annotators.

The annotated dataset[3] (03h44m) is randomly splitted into DEV and TEST subsets, with 605 dialogues (75.25%, 02h50m) for the final TEST set. The full ARCOM 2023 dataset is not public as it requires specific research agreements to be used.

[3]It can be downloaded free of charge for research purposes at `https://www.ina.fr/recherche/dataset-project` under the name `is24_news_topic`.

4. Methodology

In this section, we explain the methodology employed to construct classifiers that can predict the subject categories defined in the annotated dataset, from transcribed inputs. Our models are evaluated in Precision, Recall and F1-Score, both micro and macro averaged. We compare the predictions of the models with the human annotated TEST set. As each dialogue is annotated by two people, we weight the cost of an error with the average annotator probability of the topic. If a prediction is made for a category labeled by both annotators, it will count as 1 True-Positive, whereas it will count as 0.5 True-Positive and 0.5 False-Positive if it was annotated by only one annotator. The same principle is applied to negative cases.

4.1. Baseline BERT classification models

In the past few years, the most common approach to this kind of classification problems has been to *finetune* a language model such as BERT, with an added classification feedforward layer. We employ this technique as our first *baseline* method. We finetune three kind of French BERT-based language models : CamemBERT [23], FlauBERT [24], and FlauBERT-Oral [9]. CamemBERT and FlauBERT are pretrained with the Masked Language Modeling objective on various text sources such as Wikipedia, books, and web crawls. FlauBERT-Oral models are pre-trained on automatically transcribed texts of broadcast news. As such, we expect these models to perform better on our task, compared to the original FlauBERT models. Multiple model sizes exists: base (110M to 138M parameters) and large (335M to 373M) variants. Some models are pre-trained on text with case information, while others are pre-trained on uncased text. Models are finetuned on the DEV set of our annotated corpus. However, we randomly split this set into a training (80%) and validation (20%) subsets, so that we can optimize certain hyperparameters and monitor the performances during training without compromising the final annotated TEST set. Models are trained up to 100 epochs, with a validation every 10 model weight updates. We keep the best performing checkpoint (F1-score) on the validation set.

4.2. Mixtral-8x7B few-shot classification

As our annotated dataset is rather small, and considered only for development and evaluation purposes, we employed a few-shot classification scheme as a second classification technique. We use the Mixtral-8x7B-v0.1-*instruct* [25] language model, in conversational mode. The model is made of 47B parameters, but only uses 13B during inference. As of early 2024, it is one of the best performing open-source instruct language models in French [25]. We use a 4-bit AWQ quantization [26] of this model[4] for it to fit in our GPU cards. We prompt it to generate a JSON list of categories contained in the transcript of the dialogue. The list was provided along with the description of all categories. The outputs were post-processed to remove hallucinations, unwanted explanations and newly created classes. Three sample dialogues were included as examples in the prompt, mainly to instruct the model about the required input/output format, and not about the category usage. Those three examples belongs to three categories out of the 18 available. The full prompt along with the post-processing recipe are released with our source code[5].

[4] https://hf.co/casperhansen/mixtral-instruct-awq

[5] https://github.com/ina-foss/is24_news_topic

This model was used on the annotated TEST set to assess its performances, as well as on a random sample of unannotated dialogues. Using two NVIDIA A100 40Gb GPU cards, we were able to process 353k unannotated dialogues with an average speed of 0.58 dialogues per second.

4.3. Teacher/Student models

We explore the use of the Mixtral-generated annotations as a *finetuning* dataset for classification models. As in section 4.1, we *finetune* BERT models to classify the dialogues. We use the 353k dialogues annotated by Mixtral (considered as teacher) as the training set. Concerning the DEV dataset, used to monitor the F1-score during training, we use the 199 dialogues from the human-annotated DEV dataset. Unlike in section 4.1, the BERT models here (student) are *finetuned* for only 3 epochs, due to the higher quantity of data available for training. On a single NVIDIA RTX 4090 GPU with the *camembert-base* (CamB) model, we were able to process around 70 dialogues per second, resulting in a speedup of 120.7x over the Mixtral model which also required GPUs 9.6x more expensive.

5. Results

5.1. Automatic system evaluation

Table 2: *F1-score, Precision (P) and Recall (R) on the annotated* TEST *set for models: CamB (camembert-base), CamL (camembert-large), FlauBU (flaubert-base-uncased), FlauBC (flaubert-base-cased), FlauLC (flaubert-large-cased), FlauOF (flaubert-oral-ft), FlauOM (flaubert-oral-mixed), FlauOA (flaubert-oral-asr).*

	Model	Micro (%)			Macro (%)		
		F1	P	R	F1	P	R
Baseline	**CamB**	50.5	77.2	37.5	24.3	82.4	20.8
	CamL	58.5	73.0	48.8	37.3	82.5	31.8
	FlauLC	55.6	69.0	46.5	40.0	69.6	34.2
Mixtral-8x7B		58.6	63.0	54.8	53.8	59.8	51.5
Teacher/Student	**CamB**	**62.5**	71.7	55.3	**58.5**	68.9	53.0
	CamL	60.9	71.7	53.0	55.9	67.3	49.8
	FlauBU	60.3	69.1	53.5	54.7	64.3	50.0
	FlauBC	60.8	70.8	53.3	56.0	66.5	49.5
	FlauLC	62.2	72.9	54.3	57.3	68.7	51.4
	FlauOF	61.3	69.0	55.1	56.5	64.3	52.6
	FlauOM	62.0	74.5	53.2	54.8	67.3	48.4
	FlauOA	**62.6**	73.3	54.5	55.9	68.1	49.3

Table 2 presents the results obtained with the three types of models detailed in section 4: the finetuned baseline BERT models, the Mixtral-8x7B model in few-shot classification mode, and the Teacher/Student BERT models, finetuned on synthetic annotations generated by the Mixtral-8x7B model. Bootstrapping ($N = 1000$) confidence intervals [27] at 95% are computed. All models obtained an error margin $\leq 3.38\%$ for Micro-F1 and $\leq 3.96\%$ for Macro-F1. For the baseline BERT models, we only show the three best models, namely *camembert-large* (CamL) and *flaubert-large-cased* (FlauLC). While the few-shot prompted LLM model (Mixtral-8x7B) outperforms all baselines models, it is associated with a high computational cost (both inference time and hardware requirements). The Teacher/Student strategy was found to systematically enhance Micro-F1 scores by 2.4 to 15.0 points, and Macro-F1 scores by

17.3 to 34.2 over the baseline models.

We can notice the imbalanced aspect of the categories, with some of the best models in Micro-F1 having low macro-averaged results. Next, we see that training models on synthetic annotations in a Teacher/Student manner allows for better results than the base Teacher model (Mixtral-8x7B). All students models performed better than the base model used to generate the annotations, while having a much lower inference cost. It seems the pre-training of FlauBERT-Oral with speech transcripts of broadcast news allows for better results over the FlauBERT base model. However, FlauBERT large model, probably advantaged by its heavier architecture, obtains similar results with Oral models. The overall best results are obtained by the *camembert-base* model with a Macro-F1 of 58.5%, and a Micro-F1 equivalent to the *flaubert-oral-asr* (62.5% vs 62.6%). We therefore choose to process the whole 2.1M dialogues dataset using the *camembert-base* model. One could argue that the model used to analyze more than 10k hours of speech only obtains a 62.5% Micro-Averaged F1-score with human generated labels. While true, it is however important to keep in mind that 1) the task in itself is difficult, even for human annotators who only agree with an alpha of 0.60 ; 2) the metric used is challenging, as annotator disagreements will sanction the model whatever its predictions are ; 3) the end goal is to monitor thematic representation differences between men and women. We can conjecture the model should behave the same way on men and women transcripts. As a result, observed differences between genders should mean there is a bias in the representation of each gender depending on the subject.

5.2. Gender representation biases in broadcast news

Dialogue-level categories are mapped back to the segment-level speech utterances. We use `inaSpeechSegmenter` v0.7.7 with its default gender prediction model [28] to measure male and female speech duration of the corresponding segments. A recent evaluation of the tool by [29] on a representative dataset of TV news showed an Identification Error Rate (IER) of 6.5%.

Globally, women have an average speaking time of 36.58% (3,417h), out of 9,340 hours of speech predicted as either male or female. Figure 1.a illustrates instances where women have a lower than average speaking time, in particular in the *sport* category. Conversely, we can see a higher than average speaking time for women in some topics such as *weather*, where the speaking times parity is at 52.0%, making it the only subject where women are more involved than men. Other higher-than-average topics include *health* (47.4%), *education* (45.1%), *disaster-accident* (44.3%) and *commercial* (44.3%).

In Figure 1.b, we notice that the topic distribution is roughly similar to our manually annotated dataset (Table 1). The topic distribution here is plotted by gender, meaning that at equal speaking time, men are more likely to speak of armed issues (*unrest-conflicts-war*) than women. On the contrary, women are more likely to speak of *arts-culture-entertainment* than men.

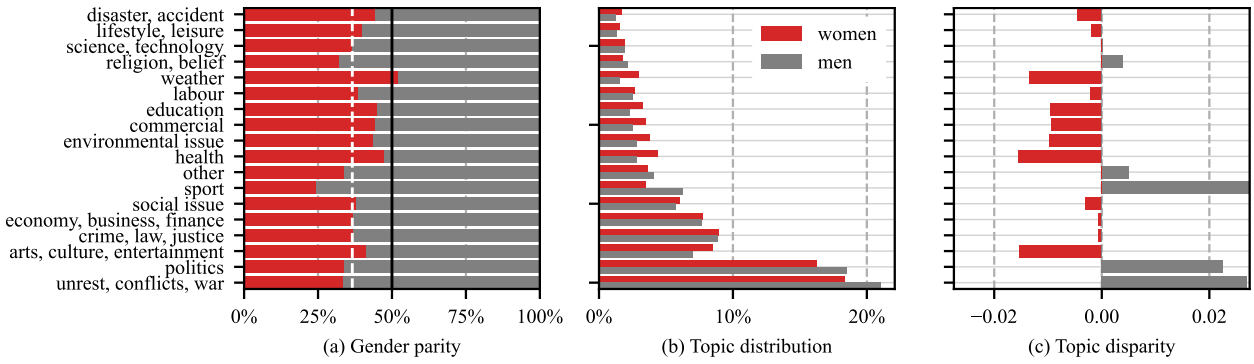
Figure 1.c plots the gender disparity, i.e. the difference between men and women topic usage relative to the global gender parity (36.58%). This figure allows for better visualization of gender differences present in Figure 1.b. A value greater than 0 means the subject is more predominantly used by men, whereas a value lower than 0 means the subject is more women-specific. It is important to remember that, except for *weather*, none of the topics achieve the gender equality.

We compute our analysis on both private and public service channels. On 9 public channels (3,678h), women have a speaking time of 40.5%, while on private channels (5,663h), women speaking time is at 34.1%. More work is required to describe the factors explaining these differences such as channel editorial, funding and human resources policies. Moreover, we show some news topics suffer from gender biases and if their distribution is not equal across channels, it may impact the speaking time differences between genders.

6. Conclusion

We focus on gender bias topic representations in French broadcast news programs. We build a corpus of topic classification in order to measure thematic differences between men and women speech. We show that certain subjects remains highly monopolized by men, such as *sports*. These results are coherent with previous similar studies, such as the CAC (*Consell de l'Audiovisual de Catalunya*) [1], NWCI (National Women's Council of Ireland) [2] and ARCOM manual thematic studies [3]. For future works, we want to explore the other annotations provided in our dataset, in particular the scope significance. The use of the annotated mixed-subjects information along with more advanced methods of topic segmentation into *dialogues* [30] could provide valuable insights, although processing more than 11k hours of speech would require more computational time. We also want to conduct an in-depth analysis of representations across public/private and radio/TV channel.

A known limitation of our analysis arises from the stereotypical binary gender categorization, without considering non-binary gender identities. To the best of our knowledge, no automatic tools currently exist for non-binary gender estimation, highlighting the importance of addressing this gap [31].



(a) Gender parity (b) Topic distribution (c) Topic disparity

Figure 1: *Measured gender representation bias per topic.*

7. Acknowledgements

This work has been partially funded by the French National Research Agency under the Gender Equality Monitor (ANR-19-CE38-0012) and Pantagruel (ANR-23-IAS1-0001) projects.

8. References

[1] D. Comas d'Argemir, "Women on tv news programmes," Tech. Rep., 2009. [Online]. Available: https://www.cac.cat/sites/default/files/2019-04/Q33_Comas_EN.pdf

[2] K. Walsh, J. Suiter, and O. O'Connor, *Hearing Women's Voices?* National Women's Council of Ireland, 2015.

[3] ARCOM, "La représentation des femmes à la télévision et à la radio - rapport sur l'exercice 2023," Tech. Rep., 2024. [Online]. Available: https://www.arcom.fr/nos-ressources/etudes-et-donnees/mediatheque/la-representation-des-femmes-la-television-et-la-radio-rapport-sur-lexercice-2023

[4] S. Macharia, "Global media monitoring project (gmmp)," *The international encyclopedia of gender, media, and communication*, pp. 1–6, 2020.

[5] S. Doumit and A. Minai, "Online news media bias analysis using an lda-nlp approach," in *International Conference on Complex Systems*, 2011.

[6] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[7] J. Cagé, N. Hervé, and B. Mazoyer, "Social media influence mainstream media: Evidence from two billion tweets," *Available at SSRN 3663899*, 2020.

[8] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022.

[9] V. Pelloin, F. Dary, N. Hervé, B. Favre, N. Camelin, A. Laurent, and L. Besacier, "ASR-Generated Text for Language Model Pre-training Applied to Speech Tasks," in *Proc. Interspeech 2022*, 2022, pp. 3453–3457.

[10] O. De Clercq, L. De Bruyne, and V. Hoste, "News topic classification as a first step towards diverse news recommendation," *Computational Linguistics in the Netherlands Journal*, vol. 10, pp. 37–55, 2020.

[11] B. Fatemi, F. Rabbi, and A. L. Opdahl, "Evaluating the effectiveness of gpt large language model for news classification in the iptc news ontology," *IEEE Access*, vol. 11, pp. 145 386–145 394, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3345414

[12] E. Leopold and J. Kindermann, "Content classification of multimedia documents using partitions of low-level features," *JVRB-Journal of Virtual Reality and Broadcasting*, vol. 3, no. 6, 2007.

[13] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, and E. Bernard, "Nuner: Entity recognition encoder pre-training via llm-annotated data," 2024.

[14] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS 2014 Deep Learning Workshop*, 2015.

[16] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "GPT3Mix: Leveraging large-scale language models for text augmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2225–2239.

[17] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4195–4205.

[18] R. Smith, J. A. Fries, B. Hancock, and S. H. Bach, "Language models in the loop: Incorporating prompting into weak supervision," *ACM / IMS J. Data Sci.*, nov 2023.

[19] W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, "UniversalNER: Targeted distillation from large language models for open named entity recognition," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=r65xfUb76p

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[21] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *Interspeech 2023*, 2023.

[22] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[23] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219.

[24] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, "Flaubert: Unsupervised language model pre-training for french," in *Proceedings of The 12th Language Resources and Evaluation Conference*, May 2020, pp. 2479–2490.

[25] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," 2024.

[26] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, C. Gan, and S. Han, "Awq: Activation-aware weight quantization for llm compression and acceleration," 2023.

[27] L. Ferrer and P. Riera, "Confidence Intervals for evaluation in machine learning." [Online]. Available: https://github.com/luferrer/ConfidenceIntervals

[28] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5214–5218.

[29] D. Doukhan, C. Maertens, W. Le Personnic, L. Speroni, and R. Dehak, "InaGVAD : A challenging French TV and radio corpus annotated for speech activity detection and speaker gender segmentation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, May 2024, pp. 8963–8974.

[30] M. Purver, *Topic Segmentation*. John Wiley & Sons, Ltd, 2011, ch. 11, pp. 291–317. [Online]. Available: https://doi.org/10.1002/9781119992691.ch11

[31] S. Ellis, S. Goetze, and H. Christensen, "Moving towards non-binary gender identification via analysis of system errors in binary gender classification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.